Code Summarization: Results Achieved with Zero-shot + Instructions

Human	**CL 7B**	**CL 13B**	**CL 34B**	**GPT 3.5**	**GPT 4**
GPT-4 (4.67)	GPT-3.5 (4.34)	GPT-4 (4.00)	GPT-4 (4.62)	CL 34B (4.15)	GPT-4 (4.10)
GPT-3.5 (4.11)	GPT-4 (4.25)	CL 34B (3.96)	GPT-3.5 (4.54)	CL 13B (4.14)	GPT-3.5 (3.81)
CL 34B (3.73)	CL 7B (4.23)	CL 13B (3.95)	CL 7B (4.52)	GPT-4 (4.03)	CL 34B (3.56)
CL 13B (3.59)	human (4.21)	GPT-3.5 (3.94)	CL 34B (4.41)	CL 7B (4.02)	CL 13B (3.36)
human (2.97)	CL 13B (4.20)	CL 7B (3.94)	CL 13B (4.28)	GPT-3.5 (3.99)	CL 7B (3.24)
CL 7B (2.89)	CL 34B (4.12)	human (3.89)	human (3.80)	human (3.87)	human (3.12)

Table 1: **(Java)** Content adequacy: ranking of the generators of summaries according to each judge, including both humans and LLMs.

Human	**CL 7B**	**CL 13B**	**CL 34B**	**GPT 3.5**	**GPT 4**
GPT-4 (4.54)	CL 7B (4.89)	CL 13B (4.31)	GPT-4 (4.54)	GPT-4 (4.25)	GPT-4 (3.99)
GPT-3.5 (3.99)	CL 34B (4.88)	CL 34B (4.22)	GPT-3.5 (4.46)	GPT-3.5 (4.06)	GPT-3.5 (3.72)
CL 13B (3.05)	CL 13B (4.88)	GPT-3.5 (4.20)	CL 7B (4.38)	CL 7B (3.97)	CL 7B (2.88)
CL 7B (2.95)	human (4.88)	CL 7B (4.18)	CL 34B (4.33)	CL 13B (3.90)	CL 13B (2.68)
human (2.93)	GPT-4 (4.87)	GPT-4 (4.16)	human (4.25)	CL 34B (3.86)	CL 34B (2.51)
CL 34B (2.78)	GPT-3.5 (4.86)	human (4.14)	CL 13B (4.10)	human (3.82)	human (2.48)

Table 2: **(Python)** Content adequacy: ranking of the generators of summaries according to each judge, including both humans and LLMs.

Code Summarization: Results Achieved with Automated Chain-of-Thought

Human	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4
GPT-4 (4.67)	GPT-4 (3.27)	GPT-4 (3.59)	CL 34B (3.95)	GPT-4 (4.97)	GPT-4 (4.63)
GPT-3.5 (4.11)	GPT-3.5 (3.26)	CL 13B (3.18)	GPT-4 (3.88)	CL 34B (4.85)	GPT-3.5 (4.04)
CL 34B (3.73)	CL 34B (3.14)	GPT-3.5 (3.16)	CL 13B (3.87)	GPT-3.5 (4.85)	CL 34B (3.78)
CL 13B (3.59)	CL 7B (3.09)	CL 34B (3.14)	GPT-3.5 (3.85)	CL 13B (4.76)	CL 13B (3.52)
human (2.97)	human (2.96)	CL 7B (3.13)	CL 7B (3.81)	CL 7B (4.72)	human (3.19)
CL 7B (2.89)	CL 13B (2.90)	human (2.95)	human (3.71)	human (4.44)	CL 7B (3.10)

Table 3: **(Java)** Content adequacy: ranking of the generators of summaries according to each judge, including both humans and LLMs.

Human	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4
GPT-4 (4.54)	human (2.55)	CL 7B (3.17)	GPT-4 (4.15)	GPT-4 (4.97)	GPT-4 (4.17)
GPT-3.5 (3.99)	CL 7B (2.53)	GPT-3.5 (3.09)	CL 7B (3.93)	GPT-3.5 (4.86)	GPT-3.5 (3.27)
CL 13B (3.05)	CL 13B (2.47)	CL 34B (2.98)	GPT-3.5 (3.87)	CL 7B (4.82)	CL 7B (2.68)
CL 7B (2.95)	GPT-3.5 (2.43)	human (2.87)	CL 13B (3.85)	CL 13B (4.72)	CL 13B (2.56)
human (2.93)	GPT-4 (2.34)	CL 13B (2.83)	human (3.85)	CL 34B (4.64)	human (2.46)
CL 34B (2.78)	CL 13B (2.05)	GPT-4 (2.65)	CL 34B (3.74)	human (4.44)	CL 34B (2.43)

Table 4: **(Python)** Content adequacy: ranking of the generators of summaries according to each judge, including both humans and LLMs.

Code Summarization: Results Achieved with Automated Chain-of-Thought + Instructions

Human	**CL 7B**	**CL 13B**	**CL 34B**	**GPT 3.5**	**GPT 4**
GPT-4 (4.67)	CL 34B (4.50)	CL 7B (3.86)	CL 13B (4.05)	GPT-4 (4.59)	GPT-4 (4.52)
GPT-3.5 (4.11)	CL 13B (4.38)	GPT-4 (3.73)	CL 34B (3.90)	CL 34B (4.41)	GPT-3.5 (4.02)
CL 34B (3.73)	GPT-4 (4.34)	CL 13B (3.71)	GPT-3.5 (3.86)	GPT-3.5 (4.34)	CL 34B (3.38)
CL 13B (3.59)	CL 7B (4.27)	CL 34B (3.55)	GPT-4 (3.85)	CL 13B (4.33)	CL 13B (3.30)
human (2.97)	GPT-3.5 (4.18)	human (3.54)	CL 7B (3.48)	CL 7B (4.25)	human (2.89)
CL 7B (2.89)	human (3.94)	GPT-3.5 (3.53)	human (3.41)	human (4.10)	CL 7B (2.68)

Table 5: **(Java)** Content adequacy: ranking of the generators of summaries according to each judge, including both humans and LLMs.

Human	**CL 7B**	**CL 13B**	**CL 34B**	**GPT 3.5**	**GPT 4**
GPT-4 (4.54)	GPT-4 (4.36)	GPT-4 (4.04)	GPT-3.5 (4.07)	GPT-4 (4.70)	GPT-4 (4.17)
GPT-3.5 (3.99)	human (4.28)	CL 13B (3.83)	CL 7B (3.99)	GPT-3.5 (4.32)	GPT-3.5 (3.09)
CL 13B (3.05)	CL 13B (4.28)	CL 7B (3.76)	GPT-4 (3.97)	CL 7B (4.27)	CL 7B (2.40)
CL 7B (2.95)	CL 34B (4.24)	CL 34B (3.73)	human (3.85)	CL 13B (4.14)	human (2.30)
human (2.93)	GPT-3.5 (4.24)	GPT-3.5 (3.67)	CL 13B (3.76)	CL 34B (4.10)	CL 13B (2.30)
CL 34B (2.78)	CL 7B (4.23)	human (3.64)	CL 34B (3.59)	human (4.05)	CL 34B (2.22)

Table 6: **(Python)** Content adequacy: ranking of the generators of summaries according to each judge, including both humans and LLMs.

Code Summarization: Results Achieved with Automated Chain-of-Thought

Human	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4
GPT-3.5 (4.91)	CL 34B (3.33)	GPT-4 (3.13)	GPT-3.5 (4.01)	CL 13B (4.06)	GPT-3.5 (4.98)
GPT-4 (4.80)	CL 7B (3.30)	human (3.01)	GPT-4 (4.00)	GPT-3.5 (4.06)	human (4.80)
human (4.80)	human (3.23)	GPT-3.5 (3.00)	CL 13B (3.99)	GPT-4 (3.98)	GPT-4 (4.73)
CL 13B (4.62)	CL 13B (3.22)	CL 13B (2.98)	CL 34B (3.97)	CL 34B (3.88)	CL 7B (4.37)
CL 34B (4.51)	GPT-3.5 (3.21)	CL 34B (2.97)	CL 7B (3.95)	CL 7B (3.77)	CL 13B (4.36)
CL 7B (4.45)	GPT-4 (3.14)	CL 7B (2.97)	human (3.94)	human (3.70)	CL 34B (4.11)

Table 7: **(Java)** Conciseness: ranking of the generators of summaries according to each judge, including both humans and LLMs.

Human	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4
GPT-4 (4.82)	GPT-4 (4.34)	GPT-3.5 (4.38)	GPT-3.5 (5.00)	GPT-4 (5.00)	GPT-4 (5.00)
CL 34B (4.81)	CL 34B (4.22)	GPT-4 (4.37)	CL 13B (4.99)	GPT-3.5 (4.98)	GPT-3.5 (4.98)
CL 13B (4.68)	GPT-3.5 (4.21)	human (4.35)	GPT-4 (4.99)	CL 34B (4.91)	CL 7B (4.87)
GPT-3.5 (4.67)	CL 7B (4.19)	CL 13B (4.31)	CL 7B (4.94)	CL 13B (4.86)	CL 13B (4.84)
CL 7B (4.40)	CL 13B (4.17)	CL 7B (4.25)	CL 34B (4.91)	CL 7B (4.82)	CL 34B (4.80)
human (3.71)	human (4.17)	CL 34B (4.07)	human (4.87)	human (4.26)	human (4.46)

Table 8: **(Java)** Fluency & Understandability: ranking of the generators of summaries according to each judge, including both humans and LLMs.

Human	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4
GPT-4 (4.54)	human (4.45)	human (4.33)	GPT-4 (4.32)	GPT-4 (4.65)	GPT-4 (4.40)
GPT-3.5 (3.99)	GPT-3.5 (4.43)	GPT-4 (4.32)	GPT-3.5 (4.15)	GPT-3.5 (4.32)	GPT-3.5 (3.94)
CL 13B (3.05)	CL 7B (4.42)	GPT-3.5 (4.31)	human (4.14)	CL 7B (4.12)	CL 7B (2.86)
CL 7B (2.95)	CL 13B (4.41)	CL 7B (4.26)	CL 34B (4.14)	CL 13B (3.90)	CL 13B (2.70)
human (2.93)	GPT-4 (4.33)	CL 13B (4.11)	CL 7B (4.13)	CL 34B (3.76)	CL 34B (2.50)
CL 34B (2.78)	CL 34B (4.33)	CL 34B (4.07)	CL 13B (3.93)	human (3.68)	human (2.46)

Table 9: **(Python)** Content adequacy: ranking of the generators of summaries according to each judge, including both humans and LLMs.

Human	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4
GPT-3.5 (4.92)	human (4.24)	human (3.71)	human (4.08)	GPT-4 (3.56)	GPT-3.5 (4.76)
human (4.91)	GPT-3.5 (4.23)	CL 34B (3.63)	GPT-4 (4.03)	GPT-3.5 (3.45)	GPT-4 (4.12)
CL 13B (4.49)	CL 7B (4.16)	CL 13B (3.59)	GPT-3.5 (3.99)	CL 7B (3.15)	human (4.11)
CL 34B (4.43)	CL 13B (4.13)	CL 7B (3.55)	CL 7B (3.91)	CL 13B (3.12)	CL 13B (3.69)
GPT-4 (4.20)	GPT-4 (3.96)	GPT-3.5 (3.55)	CL 34B (3.87)	CL 34B (2.98)	CL 34B (3.37)
CL 7B (3.95)	CL 34B (3.95)	GPT-4 (3.42)	CL 13B (3.81)	human (2.87)	CL 7B (3.37)

Table 10: **(Python)** Conciseness: ranking of the generators of summaries according to each judge, including both humans and LLMs.

Human	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4
GPT-4 (4.68)	human (4.55)	CL 7B (4.46)	CL 34B (4.18)	GPT-4 (4.80)	GPT-4 (5.00)
GPT-3.5 (4.65)	GPT-3.5 (4.51)	CL 34B (4.38)	GPT-4 (4.16)	GPT-3.5 (4.70)	GPT-3.5 (4.96)
CL 13B (4.06)	CL 13B (4.50)	CL 13B (4.35)	GPT-3.5 (4.12)	CL 7B (4.40)	CL 7B (4.14)
human (3.92)	CL 7B (4.39)	GPT-3.5 (4.34)	human (4.05)	CL 13B (4.37)	CL 13B (4.13)
CL 7B (3.89)	CL 34B (4.37)	GPT-4 (4.32)	CL 7B (4.00)	CL 34B (4.08)	human (3.93)
CL 34B (3.71)	GPT-4 (4.36)	human (4.26)	CL 13B (3.90)	human (3.92)	CL 34B (3.90)

Table 11: **(Python)** Fluency & Understandability: ranking of the generators of summaries according to each judge, including both humans and LLMs.